UNIVERSITÄT BIELEFELD
Faculty of Technology

CeBiTec
Center for Biotechnology

Hierarchical organization of syntenic blocks in large genomic datasets

Daniel Doerr

Faculty of Technology and Center for Biotechnology (CeBiTec), Bielefeld University

Introduction

Synteny hierarchies for permutations

Synteny hierarchies for sequences

PSyCHO

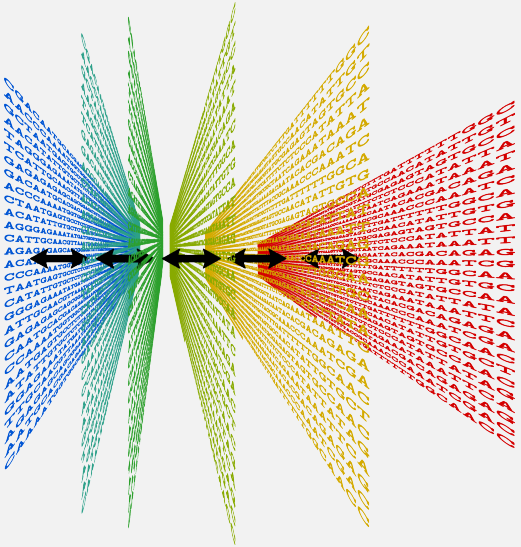
Data structures for large-scale comparisons

Objective:

- multi-species whole-genome comparisons

Solution:

- pan-genome data structures



Data structures for large-scale comparisons

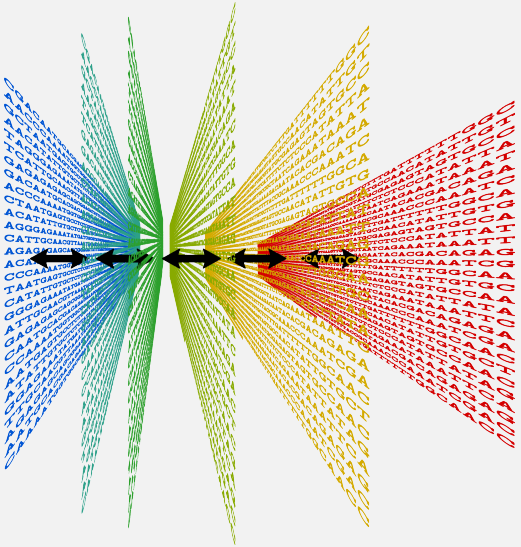
Objective:

- multi-species whole-genome comparisons

Solution:

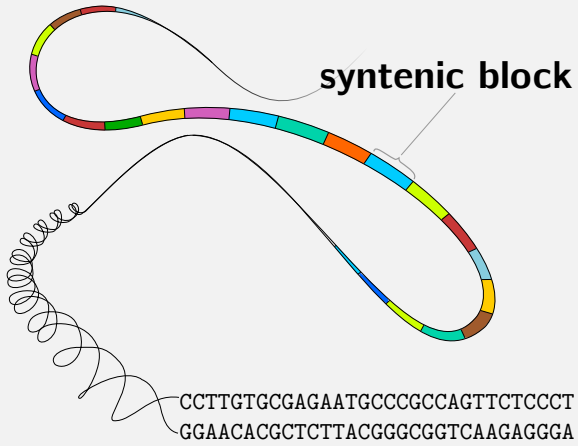
- pan-genome data structures

... only suitable for very similar genomes



Abstraction by decomposition

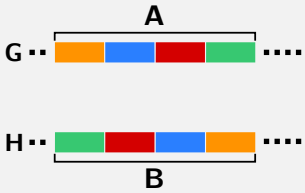
- genomes decomposed into *syntenic blocks*
- essential for studying genome evolution between distant species
  - current studies restricted to protein-coding genes
  - omission of many other conserved genomic regions



What is synteny?

A zoo of definitions:

- "*the same ribbon*" (Renwick, 1971) , set of markers co-located on same chromosome
- markers must be collinear
- local rearrangements allowed
- mostly tool-centric: FISH, GRIMM/DRIMM-Synteny, Cyntenator, i-ADHoRe, Sibelia, CoGe, Satsuma, *etc.*



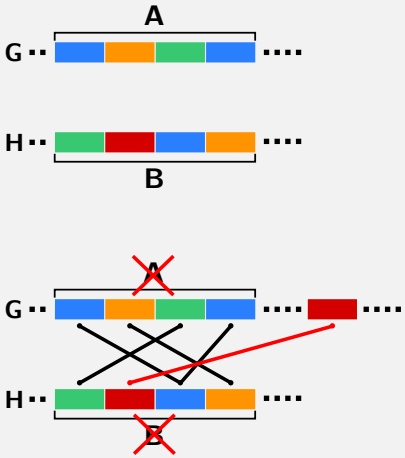
What is synteny?

homology assignment: set $\mathcal{H}$ of pairwise (equivalence) relations

Definition [Ghiurcuta and Moret, 2014]

Given two genomes G, H and homology assignment $\mathcal{H}$, two SBs $A \subseteq G$ and $B \subseteq H$ are homologous if for each

- $a \in A$: $\exists\,(a,h) \in \mathcal{H}, h \in H \implies (a,b') \in \mathcal{H}, b' \in B$
- $b \in B$: $\exists\,(b,g) \in \mathcal{H}, g \in G \implies (a',b) \in \mathcal{H}, a' \in A$



What is synteny?

homology assignment: set $\mathcal{H}$ of pairwise (equivalence) relations

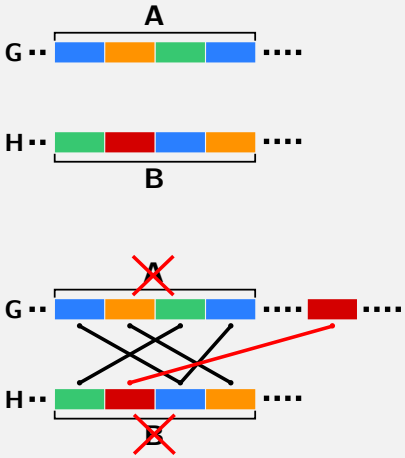
Definition [Ghiurcuta and Moret, 2014]

Given two genomes G, H and homology assignment $\mathcal{H}$, two SBs $A \subseteq G$ and $B \subseteq H$ are homologous if for each

- $a \in A$: $\exists\,(a,h) \in \mathcal{H}, h \in H \implies (a,b') \in \mathcal{H}, b' \in B$
- $b \in B$: $\exists\,(b,g) \in \mathcal{H}, g \in G \implies (a',b) \in \mathcal{H}, a' \in A$

syntenic block (SB): single marker or set of contiguous syntenic blocks

dilemma:
there is no one true decomposition of genomes into syntenic blocks

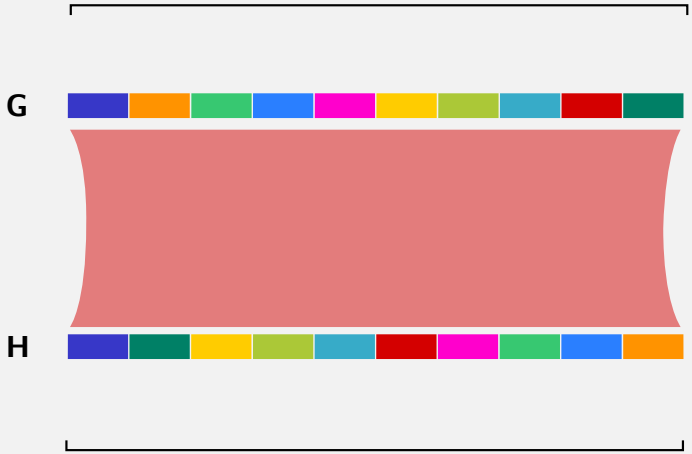


Synteny hierarchy

What are the homologous SBs of **G**,**H**?

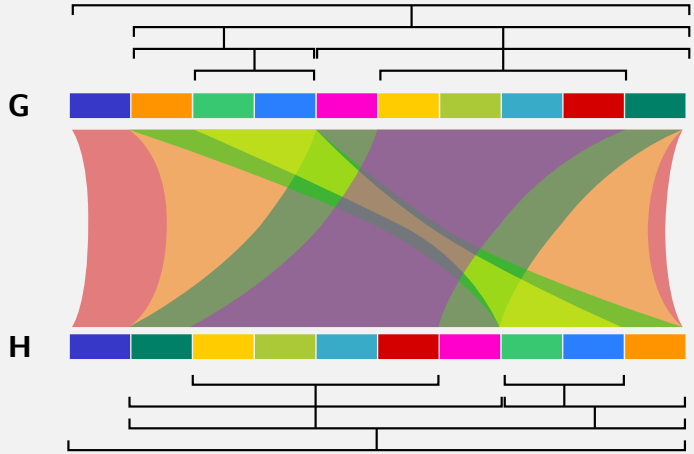
Synteny hierarchy

G, **H** are covered by one homologous SB pair



Synteny hierarchy

... but contains several other homologous SB pairs



Introduction

Synteny hierarchies for permutations

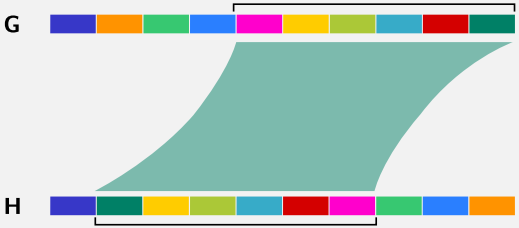
Synteny hierarchies for sequences

PSyCHO

Common intervals in permutations

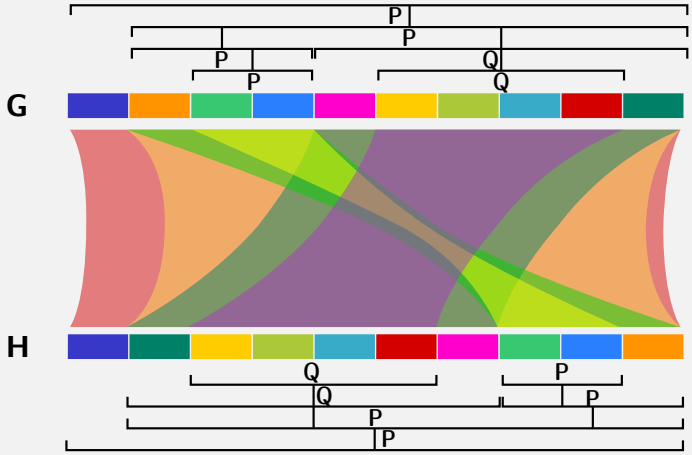
Definition

A pair of intervals of two permutations is *common* if they share the same set of elements.



Synteny hierarchy

PQ-tree: [Booth and Lueker, 1976] “**Q**”-node: collinear, “**P**”-node: permute freely



Booth and Lueker

PQ tree construction

linear time w.r.t. input size, i.e., number of 1s of an $n \times m$ matrix

- number of markers: n
- number of common intervals: $m \in O(n^2)$

... but cubic w.r.t. output size: the PQ tree has only $O(n)$ nodes!

Intervals of a PQ tree

Definition [Bergeron et al., 2008]

The *frontier* of a node is the set of labels of the leaves of the subtree rooted at this node, or a singleton comprising a leaf label.

Sets of common intervals in permutations

Definition [Bergeron et al., 2008]

A set of intervals $\mathcal{I}$ is *closed* if $(1), .., (n) \in \mathcal{I}$, $(1..m) \in \mathcal{I}$, and for each pair of intervals $(i..k), (j..l) \in \mathcal{I}$ s.t. $i < j \leq k < l$, also

$$(i..j), (j..k), (k..l), (i..l) \in \mathcal{I}$$

Sets of common intervals in permutations

Definition [Bergeron et al., 2008]

A set of intervals $\mathcal{I}$ is *closed* if $(1), .., (n) \in \mathcal{I}$, $(1..m) \in \mathcal{I}$, and for each pair of intervals $(i..k), (j..l) \in \mathcal{I}$ s.t. $i < j \leq k < l$, also

$$(i..j), (j..k), (k..l), (i..l) \in \mathcal{I}$$

Commuting sets

Definition [Bergeron et al., 2008]

Two intervals A, B commutes if

- $A \subseteq B$ or
- $B \subseteq A$ or
- $A \cap B = \emptyset$.

… and a set of intervals $\mathcal{I}$ is *commuting* if all pairs of intervals commute.

Strong intervals

Definition [Bergeron et al., 2008]

Given a set of intervals $\mathcal{I}$, an interval A is *strong* if it commutes with all intervals $B \in \mathcal{I}$.

The strong intervals of a closed set of intervals $\mathcal{I}$ are the frontier of the PQ tree of $\mathcal{I}$.

Introduction

Synteny hierarchies for permutations

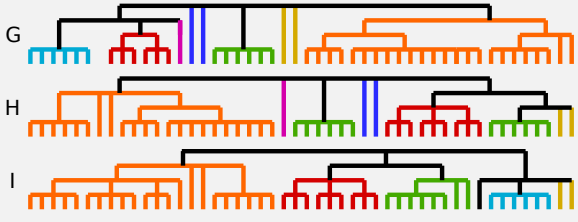
Synteny hierarchies for sequences

PSyCHO

SB hierarchy

Context-dependency

two sets of common intervals intersect *only* if all their intervals intersect in the corresponding sequences



Sets of common intervals in sequences

Definition

A set of intervals $\mathcal{I}$ is *near-closed* if $(1), .., (n) \in \mathcal{I}$, $(1..m) \in \mathcal{I}$, and for each pair of intervals $(i..k), (j..l) \in \mathcal{I}$ s.t. $i < j \leq k < l$, also

$$(i..l) \in \mathcal{I}$$

Sets of common intervals in sequences

Definition

A set of intervals $\mathcal{I}$ is *near-closed* if $(1), .., (n) \in \mathcal{I}$, $(1..m) \in \mathcal{I}$, and for each pair of intervals $(i..k), (j..l) \in \mathcal{I}$ s.t. $i < j \leq k < l$, also

$$(i..l) \in \mathcal{I}$$

Lemma

Let $\mathcal{I}$ be a near-closed set of intervals. Then there exists a unique *PQ*-tree with frontier $\mathcal{F}$ such that for the set of strong intervals $I' \subseteq I$ holds true that $\mathcal{I}' \subseteq \mathcal{F}$ and $|\mathcal{I}| \geq \lceil 1/2 \cdot |\mathcal{F}| \rceil$.

Introduction

Synteny hierarchies for permutations

Synteny hierarchies for sequences

PSyCHO

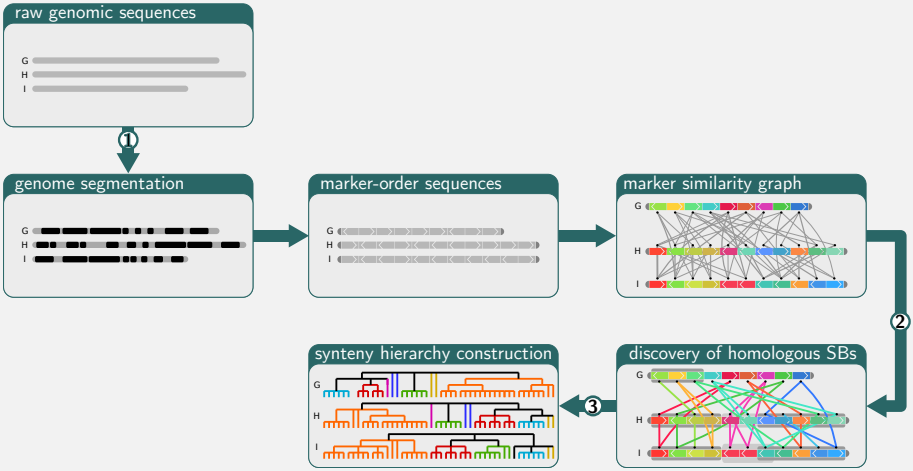
PSyCHO

PSyCHO

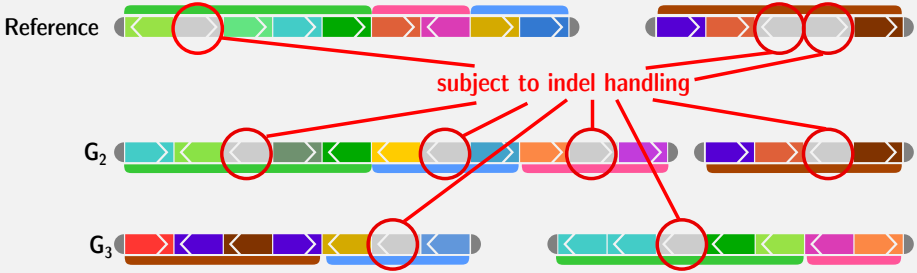
Principled Synteny using Common Intervals and Hierarchical Organization

`http://github.com/danydoerr/PSyCHO`

Construction of a synteny hierarchy

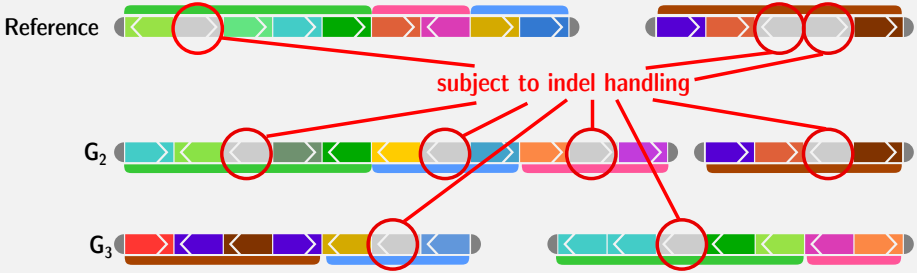


Similarity graph, syntenic contexts, homologous SBs



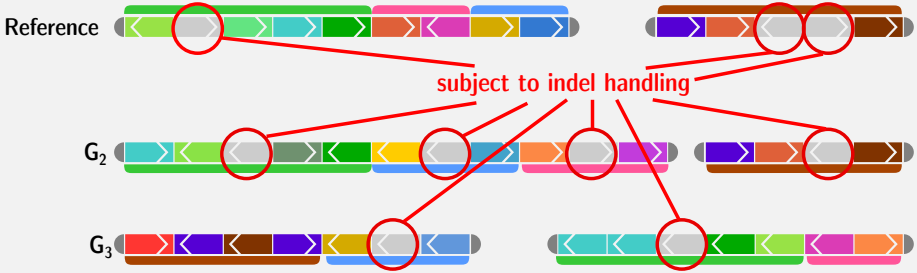
1. reference-based reconstruction of syntenic contexts
   - computational problem: finding δ-teams in sequences

Similarity graph, syntenic contexts, homologous SBs



1. reference-based reconstruction of syntenic contexts
   - computational problem: finding δ-teams in sequences
2. handling of insertions/deletions *(work in progress)*

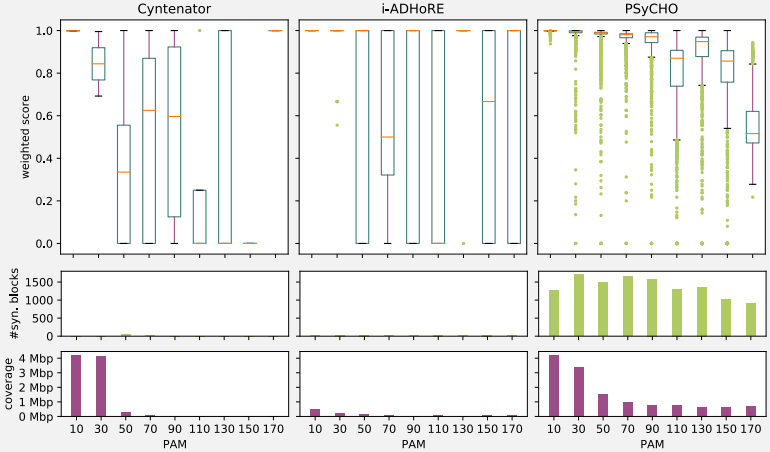
Similarity graph, syntenic contexts, homologous SBs



1. reference-based reconstruction of syntenic contexts
   - computational problem: finding δ-teams in sequences
2. handling of insertions/deletions *(work in progress)*
3. reference-based discovery of homologous syntenic blocks in each context
   - computational problem: enumerating common intervals in k sequences

Analysis of simulated genomes

5 species, 1000 markers of length 300,
point mutations+rearrangements+ins+del+dupl

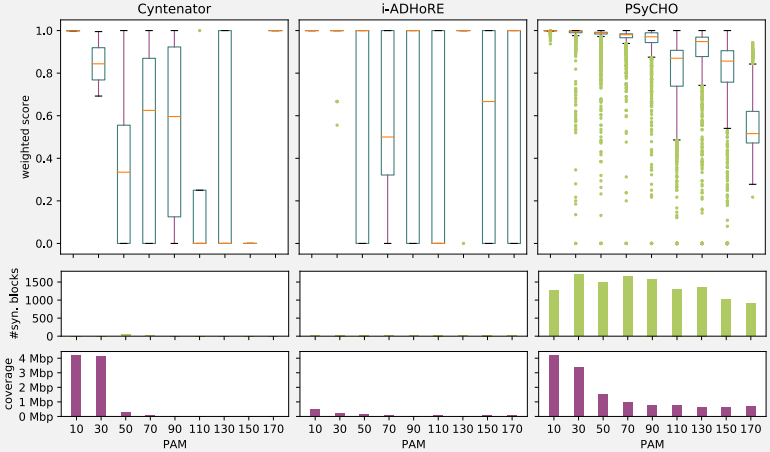


Analysis of simulated genomes

Weighted Synteny Score: Fraction of markers in a homologous set of syntenic blocks that have at least one homologous counterpart in each block or have no homologous counterpart at all in the respective genomes.

Analysis of simulated genomes

5 species, 1000 markers of length 300,
point mutations+rearrangements+ins+del+dupl



Analysis of Drosophila genomes

species	ID	scaffolds	size (Mbp)	CDSs	markers
D. *melanogaster*	D.mel	7	120.3	30,443	98,214
D. *simulans*	D.sim	6	118.2	24,119	100,549
D. *yakuba*	D.yak	6	119.5	23,304	100,774

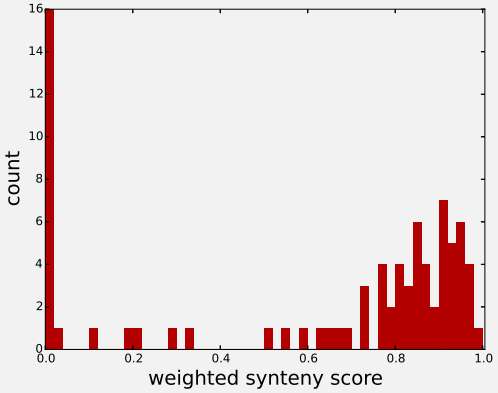
UNIVERSITÄT BIELEFELD

Analysis of Drosophila genomes

	genome	PSyCHO	i-ADHoRe
coverage	D.mel	**0.782**	0.682
	D.mel	0.823	**0.840**
	D.yak	**0.783**	0.763
#SBs		top: 10 int. nodes: 2090	80

Weighted Synteny Score [Ghiurcuta and Moret, 2014]

i-ADHoRe



PSyCHO: 1 (by def.)

Thank you!

UNIVERSITÄT BIELEFELD

References

Bergeron, A., Chauve, C., de Montgolfier, F., and Raffinot, M. (2008).
Computing common intervals of K permutations, with applications to modular decomposition of graphs.
SIAM Journal on Discrete Mathematics, 22(3):1022–1039.

Booth, K. S. and Lueker, G. S. (1976).
Testing for the Consecutive Ones Property, Interval Graphs, and Graph Planarity Using PQ-Tree Algorithms.
JCSS.

Brejová, B., Burger, M., and Vinař, T. (2011).
Automated Segmentation of DNA Sequences with Complex Evolutionary Histories.
In *Proc. of WABI 2011*, volume 6833, pages 1–13.

Ghiurcuta, C. G. and Moret, B. M. E. (2014).
Evaluating synteny for improved comparative studies.
Bioinformatics, 30(12):i9–18.

Meidanis, J. and Munuera, E. G. (1996).
A theory for the consecutive ones property.
Proceedings of WSP.

Visnovská, M., Vinař, T., and Brejová, B. (2013).
DNA Sequence Segmentation Based on Local Similarity.
In *Proc. of ITAT*, volume 1003, pages 36–43.